PLANT SCIENCE TODAY
Vol 12(sp1): 01–12
https://doi.org/10.14719/pst.9940

eISSN 2348-1900

RESEARCH ARTICLE

Predicting tomato yield under heat stress in Tamil Nadu using Machine Learning Models

Musierose C[1], Maragatham N[2]*, Sathyamoorthy N K[1], Djanaguiraman M[3], Indu Rani C[4], Somasundaram E[5], Sakthivel N[6] & Sivasakthivelan P[6]

[1]Agro Climate Research Centre, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[2]Centre for Students Welfare, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[3]Department of Crop Physiology, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[4]Department of Vegetable Sciences, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[5]Directorate of Agri Business Management, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[6]Agricultural Research Station, Tamil Nadu Agricultural University, Bhavanisagar 638 451, Tamil Nadu, India

*Correspondence email - mm65@tnau.ac.in

Received: 09 June 2025; Accepted: 21 July 2025; Available online: Version 1.0: 09 September 2025

Cite this article: Musierose C, Maragatham N, Sathyamoorthy NK, Djanaguiraman M, Indu RC, Somasundaram E, Sakthivel N, Sivasakthivelan P. Predicting tomato yield under heat stress in Tamil Nadu using Machine Learning Models. Plant Science Today. 2025;12(sp1):01–12. https:/doi.org/10.14719/pst.9940

Abstract

Rise in temperature and its unpredictability has an adverse effect on growth and yield, making it an important variable in tomato (*Solanum lycopersicum* L.) production. This study aimed at evaluating the impact of temperature variability on tomato yield and developing predictive models using Machine Learning (ML) techniques to forecast future productivity under changing climate. The tomato yield was predicted using Machine Learning Models (MLM) such as Random Forest (RF), XGBoost (XG) and K-Nearest Neighbours (KNN) in response to temperature changes. The model was evaluated and improved by comparing both Train-Test split (T-T) and K-fold cross validation techniques. Among these, the T-T method performed better and was used for model training and testing. The findings showed that RF model outperformed the others, with the T-T dataset, achieving Coefficient of Determination (R^2) = 0.84, Mean Squared Error (MSE) = 7.88, Root Mean Square Error (RMSE) = 2.81 and Mean Absolute Error (MAE) = 1.19, followed by XGBoost and KNN. Additionally, Kernel Density Estimation (KDE) correlation analysis was employed to examine the relationship between yield and temperature. Moreover, future tomato yields were predicted under Shared Socio-economic Pathways (SSP2-4.5 and SSP5-8.5) for the period of 2023-2026 using the RF model. Tomato productivity is likely to increase gradually in the immediate future and eventually fall under extreme heat. These findings illustrate the potential of machine learning in forecasting tomato yield under varying temperature conditions, thereby aiding climate adaptation strategies and agricultural planning.

Keywords: k-nearest neighbour; machine learning models; random forest; temperature; tomato; XGBoost; yield prediction

Introduction

Tomato (*Solanum lycopersicum* L.), a member of the family Solanaceae, is a commercially and nutritionally important vegetable crop cultivated worldwide, with a global production value exceeding USD 182 billion (1). It is rich in a potent antioxidant, lycopene which is an anticarcinogen (2). In India, it is grown in both tropical and sub-tropical areas and ranked second in vegetable production next to potato (3). Tamil Nadu, with its diverse agro-climatic zones, supports extensive tomato cultivation, producing 7.94 lakh tonnes annually from an area of 41392 hectares, with an average yield of 30.51 t/ha. Tomatoes are also valuable sources of ascorbic acid, carotenoids, vitamins, minerals and organic acids (4). The fruit contains several bioactive compounds associated with anti-inflammatory, antiallergenic, antibacterial, vasodilatory, antithrombotic, cardioprotective and antioxidant effects (5).

Tomato production and fruit quality are increasingly vulnerable to changing climatic variables, such as temperature, rainfall and extreme weather events. Among these, temperature is a critical factor influencing plant development and yield. Several studies have highlighted the sensitivity of tomato fruit metabolism and quality to elevated temperatures, which can lead to significant reductions in productivity (6).

The optimum temperature for tomato crop growth is 25 °C - 30 °C at daytime and 20 °C at night (7). However, this range can vary slightly depending on the location and cultivar. A rise in temperature above this threshold might induce flower abscission, reduced pollen quality, abnormal pollen, decreased fruit set percentage and eventually reduced crop yield (8). The susceptibility of the pollen grain determines the behaviour of the crop under heat stress which lowers fruit yield. Higher temperatures can significantly affect the duration between flowering and fruit formation. They alter the ratio of sugars and

acids, significantly decreasing fruit set, fruit quantity, weight and quality (carotenoids, lycopene). Prolonged heat stress (mean 34 °C/19 °C) led to 34 % flower drop resulting in a 71 % reduction in fruit set (9).

When exposed to a continuous period of daily daytime temperatures over 29 °C and nighttime temperatures above 21 °C, tomato plants lose their ability to reproduce, whereas when the daytime temperature exceeds 35 °C during flowering and fruit set, severe reproductive damage and yield loss occur (10). Daily temperatures above 40 °C could result in flower loss (11). By the end of the twenty-first century, the Intergovernmental Panel on Climate Change predicts that the average global temperature will rise by 2.6 °C - 4.8 °C. Rising temperatures also raised the crop water demand which resulted in reduced crop productivity (12). At extreme levels, considerable loss in crop production was recorded (13). Some studies have aimed at quantifying the correlation between crop production and temperature to analyse the possibility of predicting future tomato production (14). These findings highlight the need for predictive tools that can assess the potential yield losses under future climate scenarios.

Recently, ML has served as a prominent tool in predicting crop yield, demonstrating promising results in agriculture (15). ML algorithms extract patterns and relationships from datasets to make informed predictions. When trying to create a high-performance prediction model, ML research presents several possibilities and challenges. The selection of appropriate algorithms is critical for addressing the issue. These algorithms must also be capable of handling large datasets efficiently. ML-based crop yield prediction is an important area of research as it aids in decision-making by predicting future productivity. Three MLMs, RF Regressor, XGBoost Regressor and KNN Regressor were trained and tested in this study with maximum, minimum and average temperature as inputs and tomato crop yield as output data for Tamil Nadu. The MLMs were validated using performance metrics such as MSE and RMSE to determine the most accurate model for yield prediction. This study investigates the effectiveness of these models in forecasting tomato yield in Tamil Nadu based on temperature variations.

The study was conducted with the following objectives:

1. To analyze the relationship between maximum, minimum and average temperatures in Tamil Nadu during the study period.
2. To assess the relationship between temperature and tomato productivity.
3. To train, test and validate machine learning models to identify the most accurate model for predicting tomato yield in Tamil Nadu.
4. To forecast future tomato yield using the best-performing ML model and projected temperature data.

These objectives aim to provide insight into the viability of ML as a tool for predictive agriculture under future climate scenarios.

Materials and Methods

Study Area

The study was conducted in Tamil Nadu, a southern state of India. The state extends from 8°4 N—13°35 N latitude and from 76°18 E—80°20 E longitude. The state covers 130058 sq. km having a long coastline along the Bay of Bengal in the east and the Western Ghats in the west. Kerala, Karnataka and Andhra Pradesh border the state in the northwest and the Indian Ocean in the south. The average rainfall for Tamil Nadu is 945 mm. The northeast monsoon contributes 48 % and the southwest monsoon accounts for about 32 % of the total rainfall of the state. Agriculture is the major occupation of the state, having 6.26 million hectares of total cultivated area as per 2022-23 data. There are four major soil types in Tamil Nadu classified by the Tamil Nadu Agricultural University, which include red soils (62 %), black soils (12 %), laterite soils (3 %) and coastal saline soils (7 %). The state is divided into 38 districts, of which 36 districts were included in the study excluding Chennai and Nilgiris based on the area under cultivation of tomato. The total area under tomato cultivation was 41392 hectares in Tamil Nadu with a production of 794330 metric tonnes in 2022-23. The average yield of the crop was 19.2 t/ha for Tamil Nadu. Tiruvannamalai, Salem, Dharmapuri, Coimbatore, Erode, Trichy, Madurai and Dindigul are the major tomato producing districts.

Data Collection

Crop production and temperature data

Tomato crop is grown in almost every district of Tamil Nadu except Chennai and Nilgiris and is cultivated in two seasons. However, due to the unavailability of season-wise data, only the *Kharif* season (June-October) was considered for the study. The tomato crop yield data was collected from the Department of Horticulture and Plantation Crops, Chennai, Tamil Nadu, for the years 2006-2022. The temperature data was obtained from India Meteorological Department (IMD) from 1998-2022 to analyze its relationship between maximum, minimum and average temperatures in Tamil Nadu. However, temperature data from 2006-2022 was used for model validation and yield prediction due to the limited availability of yield data. Monthly mean temperatures were used, as the temperature requirement of crop varies across growth stages.

Temperature correlation analysis

Correlation analysis was performed with maximum, minimum and average temperature data using Pearson's correlation coefficient and their relationship was established to check internal temperature relationships (16).

Relationship between temperature and tomato Yield

The relationship between temperature and tomato yield was generated using the Seaborn bivariate KDE tool, which revealed the relation between monthly temperatures and crop yield.

ML models for yield prediction

Data pre-processing: Data pre-processing is the process of producing a clean dataset from raw data by handling missing values and null values. Subsequently, the data was split into training and testing sets in an 80:20 ratio, with 80 % used for model training. The ML algorithms learned from the training dataset to accurately predict future data. A larger training dataset produces more accurate results. Fig. 1 illustrates the methodology of data collection, processing, preparation model training and evaluation.

Dataset Description: Fig. 2 shows the structure of the dataset used for training and testing the MLMs. Columns 6_max_temp-

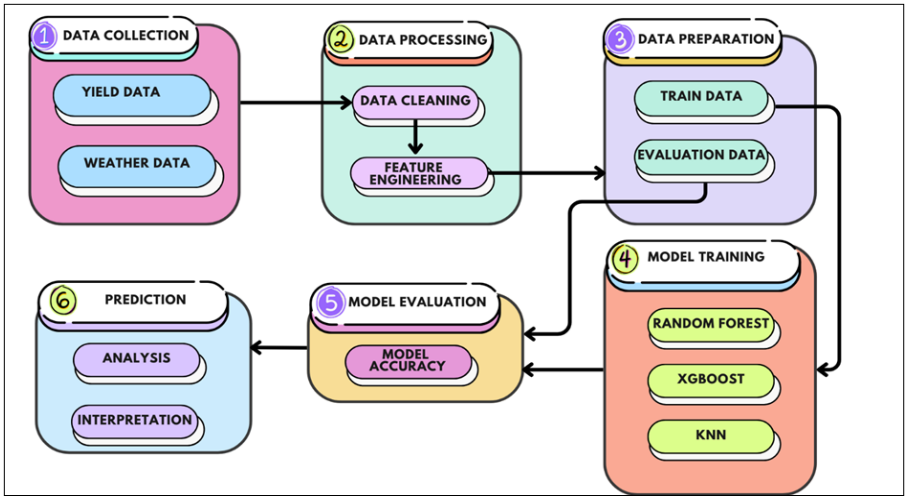

Fig. 1. Methodology of data collection, processing, model evaluation and prediction in this study.

	year	6_max_temp	7_max_temp	8_max_temp	9_max_temp	10_max_temp	6_min_temp	7_min_temp	8_min_temp	9_min_temp	10_min_temp	area	yield
count	629.00	629.00	629.00	629.00	629.00	629.00	629.00	629.00	629.00	629.00	629.00	629.00	0.63
mean	2014.00	34.42	33.48	33.05	32.84	31.81	24.77	24.32	24.01	23.78	23.19	735.77	11.94
std	4.90	2.52	2.40	2.19	1.81	1.41	1.89	1.82	1.71	1.63	1.52	1722.30	7.01
min	2006.00	26.69	26.15	26.54	25.98	26.61	18.60	18.25	18.19	17.81	17.31	0.00	0.00
25%	2010.00	32.82	32.14	31.91	31.89	31.07	23.72	23.31	23.07	22.86	22.29	1.00	8.95
50%	2014.00	34.58	33.75	33.43	33.16	32.08	25.01	24.66	24.46	24.22	23.58	50.00	13.09
75%	2018.00	36.34	35.40	34.76	34.14	32.73	26.29	25.77	25.35	25.03	24.37	448.00	15.00
max	2022.00	39.16	37.28	36.20	35.58	34.92	27.91	26.90	26.37	26.15	25.55	13542.00	39.56

Fig. 2. Dataset description for Tomato yield prediction using MLM.

10_max_temp are the mean monthly maximum temperatures from the first to the fifth month of the season (*Kharif* season) in degree Celsius (°C). The objective of this study was to assess the impact of temperature changes on tomato productivity and to examine the potential of temperature data in yield forecasting. Hence, temperature was specifically employed as the key input variable and the other weather variables were excluded. The 'area' and 'year' columns represents the area under tomato cultivation and the respective years (2006-2022). The 'yield' column provides the average seasonal yield of tomato in tonnes per hectare (t/ha). A total of 7548 observations (17 years × 37 districts × 12 variables) were used for training and testing the MLMs.

Machine Learning Models

The XGBoost regressor, RF regressor and KNN regressor are the MLMs used and evaluated in this study using Python programming.

Random Forest Regressor: RF Regressor is an ensemble learning method used for regression tasks, which constructs multiple decision trees and aggregates their outputs. It is based on statistical learning theory and decision trees (17). 'M' represents the number of decision trees selected and 'm' denotes the number of variables that are randomly selected at each node to split the node for constructing a single decision tree (Equation 1). A comprehensive regression tree was generated by repeating these steps and averaging the outputs to obtain the final prediction (18). Fig. 3 depicts the single tree used by RF Regressor for training in this study.

$$\hat{y}_i = \frac{1}{M}\sum_{m=1}^{M} f_m(x_i) \qquad (1)$$

where,

$\hat{y}_i$ = Predicted value for input x_i

M = number of decision trees

$f_m(x_i)$ = prediction from m-th decision tree

XGBoost Regressor: XGBoost was an improved version of the Gradient Boosted Decision Tree (GBDT) (19). It utilizes second order derivatives and regular terms, which improve the algorithm performance in training and quick computing. Each new tree corrects the errors of previous trees. Fig. 4 shows the single XGBoost tree used for training the model in this study.

XGBoost basic function (Equation 2): XGBoost prediction is performed by the ensemble of decision trees.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \qquad (2)$$

where,

$\hat{y}_i$ = Predicted value for input x_i

K = Total number of trees

$f_k(x_i)$ = Prediction from the k-th tree

XGBoost objective function (Equation 3):

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (3)$$

where,

$\hat{y}_i$ = Predicted value for input x_i

$l(y_i, \hat{y}_i)$ = Loss function

$\Omega(f_k)$ = Regularization term to penalize tree complexity

K- Nearest Neighbours Regression: Applying a distance metric for each pair of samples, KNN regression determines the K samples that are adjacent to each other in the training set (Equation 4). Predictions are then made using the KNN. Fig. 5 illustrates the mechanism of KNN regressor.

$$\hat{y}_i = \frac{1}{K}\sum_{j \in \mathcal{N}_{\mathcal{K}}(x_i)} y_j \qquad (4)$$

where,

$\hat{y}_i$ = Predicted value for input x_i

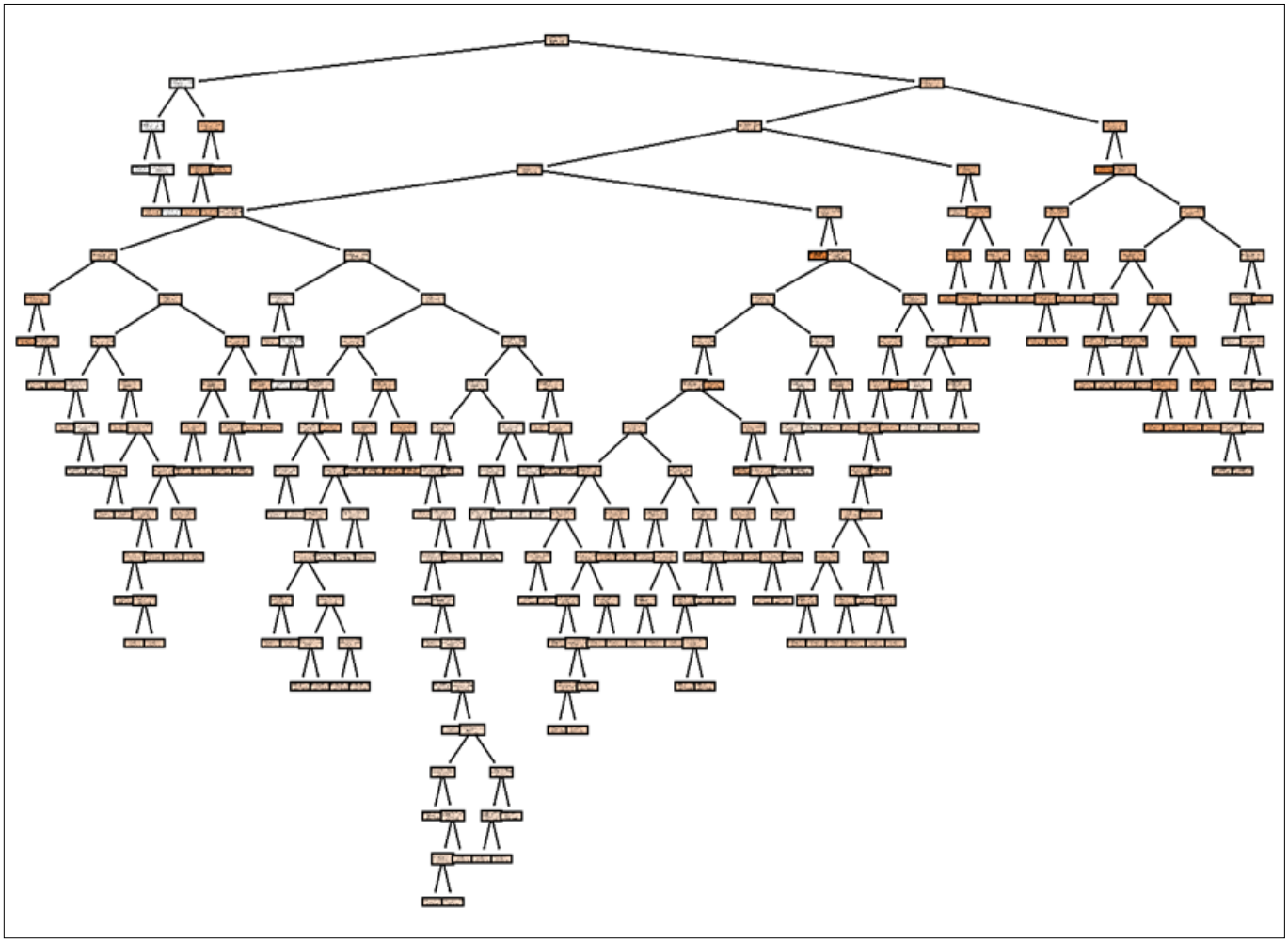

Fig. 3. A Single Random Forest Tree developed in the study.

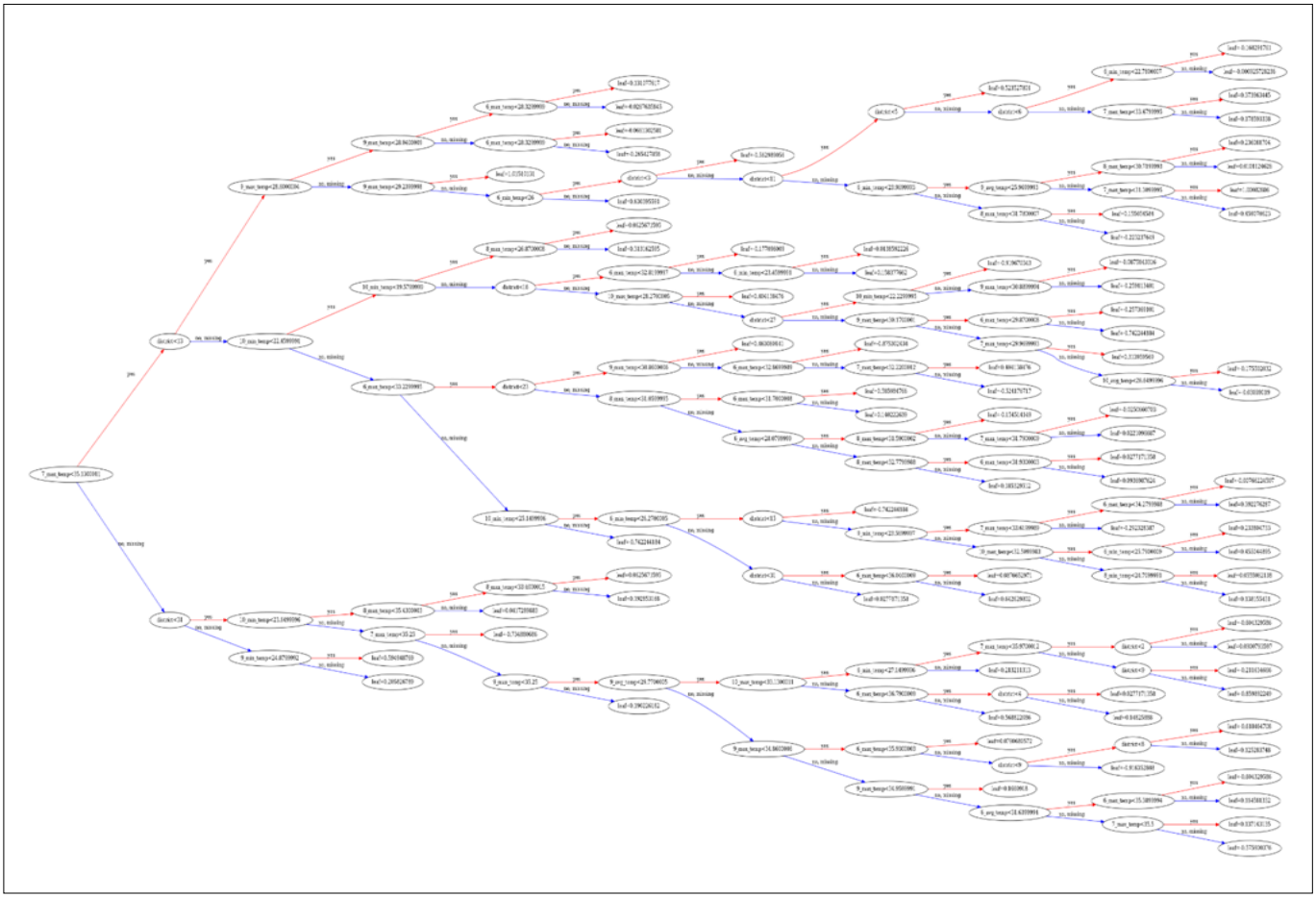

Fig. 4. A Single tree in XGBoost regressor model used in this study.

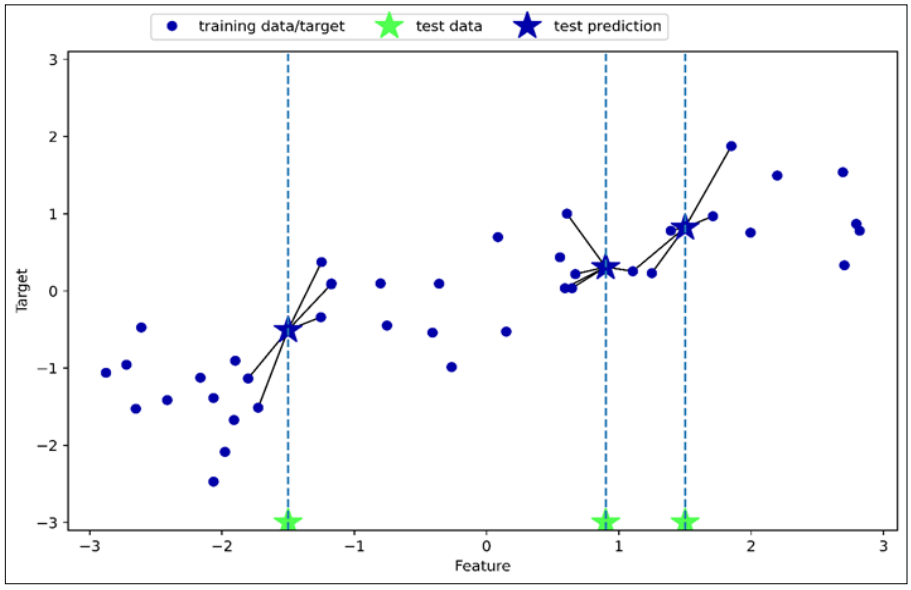

Fig. 5. KNN Regressor.

K = number of nearest neighbours

$N_K x_i$ = set of the K nearest neighbours of x_i

Y_j = actual values of the j^{th} nearest neighbours

Hyperparameters Tuning

Hyperparameter tuning was carried out using Optuna, a specialized python library known for its advanced and quick optimization methods. It searched for the best parameters by making random guesses and learning from past results, rather than trying all possible combinations. The search was refined continuously to quickly find the best values which yielded the optimal values for training the models and achieving good prediction results. Hyperparameters tuned for RF includes, max_depth (None), min_samples_leaf (1), min_samples_split (2) and n_estimators (200). The parameters such as n_neighbors (5), weights (distance), metric (minkowski) and p (2) were tuned for KNN model. The parameters including learning_rate (0.08), max_depth (10), n_estimators (471) and subsample (0.65) were tuned for XG model.

Train-test Split (T-T) and K-fold Cross Validation

In T-T, the dataset was split into two parts (80 % to train and 20 % to test) before model training. After experimenting with a variety of ratios, beginning with 50 % each and reducing the test portion, a ratio of 80 % for training and 20 % for testing was determined to be effective. The train-test split prevents overfitting and ensures better generalization of data.

To assure the accuracy of the training results, K-fold was employed, in which different k values were tested to determine K-Value with the best prediction outcomes. k = 5 was selected, as it is less skewed and has a lower cost of computation. The difference between T-T and K-fold is clearly illustrated in Fig. 6.

Model Evaluation

Validation skill scores such as R^2, MSE, RMSE and MAE were employed for evaluating the model accuracy (20). The lower the MSE, RMSE, MAE and MAPE values, the greater the model prediction accuracy. R^2 denotes the degree of fit between predicted and observed values of the model ranging from 0-1. When the R^2 is close to 1, the model is a good fit. The equations of the validation scores are as follows. MSE, RMSE, MAE and R^2 are calculated using the following formulas (equations 5 - 9).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

Where, y_i is the model predicted value, is the observed value and n is the number of data points.

Future Prediction

The MLM with the most accurate result was used for predicting future tomato yield for the successive 4 years (2023-2026). Future temperature data were collected from CMIP6 climate projections. Out of five SSP scenarios, SSP2 4.5 and SSP5 8.5 were considered for prediction, as they represent the gradual and extreme rise in temperature and change in climate.

Results

Pearson's correlation of maximum and minimum with average temperature

The relationship of maximum and minimum temperatures with the average temperature from 1998-2022 is illustrated in Fig. 7-8. Fig. 7(a-b) shows that the minimum temperature exhibited greater variability than the maximum temperature. The correlation matrix for the monthly minimum, maximum and average temperature for the study period was generated and is in the Fig. 8. The results indicated a strong positive correlation between average temperature and both minimum (R^2 = 0.90) and maximum (R^2 = 0.92) temperatures, with minimum temperature displaying slightly more influence on the average values.

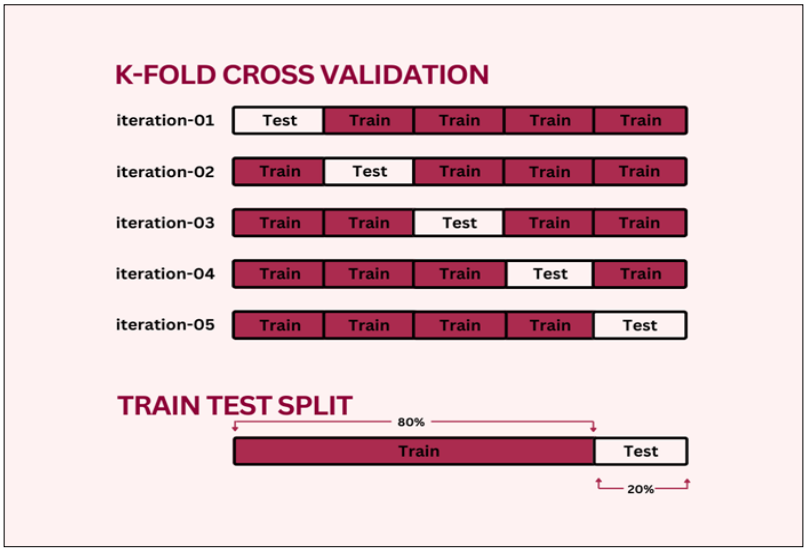

Fig. 6. Train-Test split and K-Fold Cross Validation.

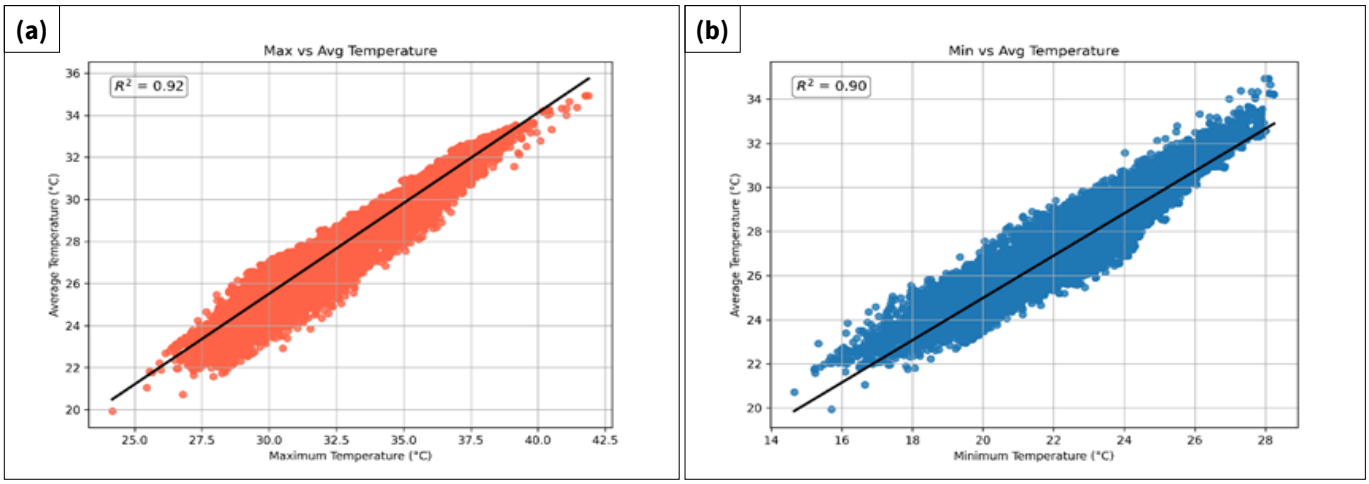

Fig. 7. Correlation of monthly average temperature with maximum (a) and minimum (b) temperature. The R^2 values were calculated for both a and b.

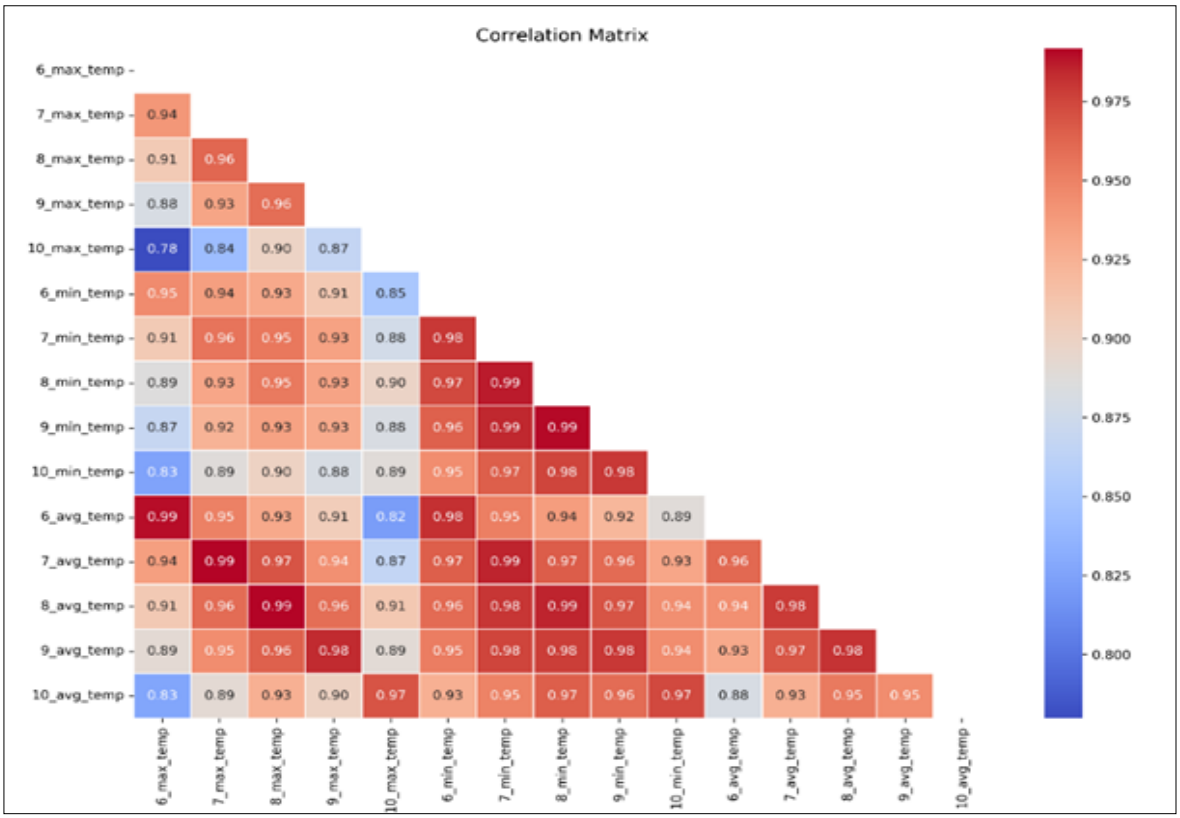

Fig. 8. Correlation matrix among the monthly maximum, minimum and average temperatures during 1998 to 2022 in Tamil Nadu; 6 - June; 7 - July; 8 - August; 9 - September; 10 - October months.

Relationship between temperature and tomato productivity

The Fig. 9 a shows the area and production of tomato in Tamil Nadu from 2006-2022. The trend indicates that tomato was cultivated for around 22000 ha until 2017, after which there was a sudden increase in the area and production of tomato in Tamil Nadu. With a few exceptions, a rise in maximum temperature resulted in a decline in tomato yield in the past years (Fig. 9 b).

The seaborn bivariate kernel density tool was adopted to generate a relationship between monthly average temperature and tomato yield across Tamil Nadu for the period of 2006-2022 (17 years).

The Kernel Density Estimation (KDE) plots in Fig. 10 (a-e) illustrate the relationship between tomato productivity and monthly average temperatures across different months of the season. A sharp increase in productivity was observed when temperature ranged between 24 °C and 27.5 °C, indicating optimal growing conditions. However, beyond 28 °C, the density of higher yield observations declined steadily. The tomato yields began to drop when average temperatures exceeded 28.5 °C, with the highest density of yields concentrated below this threshold (Fig. 10 b - c). At temperatures above 30 °C, yield values clustered in the lower range (<15 t/ha), suggesting a negative impact of heat stress. These results confirm that tomato productivity is sensitive to rising temperatures and highlight a thermal threshold beyond which yield losses become significant.

Fig. 10 (a-b) show that during the first two months of the crop season (June and July), an average temperature range of 28 °C-32 °C typically resulted in a yield of around 15 t/ha. The third and fourth months (8_avg_temp and 9_avg_temp) (Fig. 10 c-d) with an average temperature of 28 °C-30 °C produced an average yield of 12 t/ha. In the fifth month (10_avg_temp) (Fig. 10 e), the average temperature was 28 °C producing an average yield of 12 -15 t/ha. However, across all months, when the temperature was within the optimal range of 24 °C-26 °C, the maximum yield reached up to 40 t/ha, highlighting the critical role of temperature thresholds in tomato productivity. This might be due to the enhanced flowering and fruit set by maintaining pollen germination rates and pollen tube growth, which are crucial for successful fertilization and yield when the temperature was optimum (24 °C-26 °C).

Model prediction accuracy

Three machine learning models were evaluated in this study, such as RF Regressor, XGBoost Regressor and KNN Regressor. T-T split and K-fold with k = 5 were used for comparison and better results. The model performance was evaluated using four validation skill scores namely MSE, RMSE, MAE and R^2. The performance results are presented in Table 1 and Fig. 11.

RF performed better for the data than the other two models, with higher R^2 values of 0.84 and 0.82 in T-T and K-fold respectively. It was followed by XGBoost with R^2 values of 0.82 and 0.80 in T-T and K-fold respectively. The KNN performed significantly worse than RF and XG with R^2 values of 0.62 and 0.56 with T-T and K-fold.

The MSE values for the RF, XG and KNN models were 7.88, 8.76 and 18.37 with T-T and 9.12, 9.62 and 21.21 with K-fold, respectively. The RMSE scores of RF, XG and KNN were 2.81, 2.96 and 4.29 with T-T and 2.95, 3.07 and 4.60 with K-fold, respectively. Similarly, the MAE of RF, XG and KNN were 1.19, 1.60, 3.05 with T-T and 1.32, 1.75 and 3.15 with K-fold, respectively. The lower MSE, RMSE and MAE values of RF

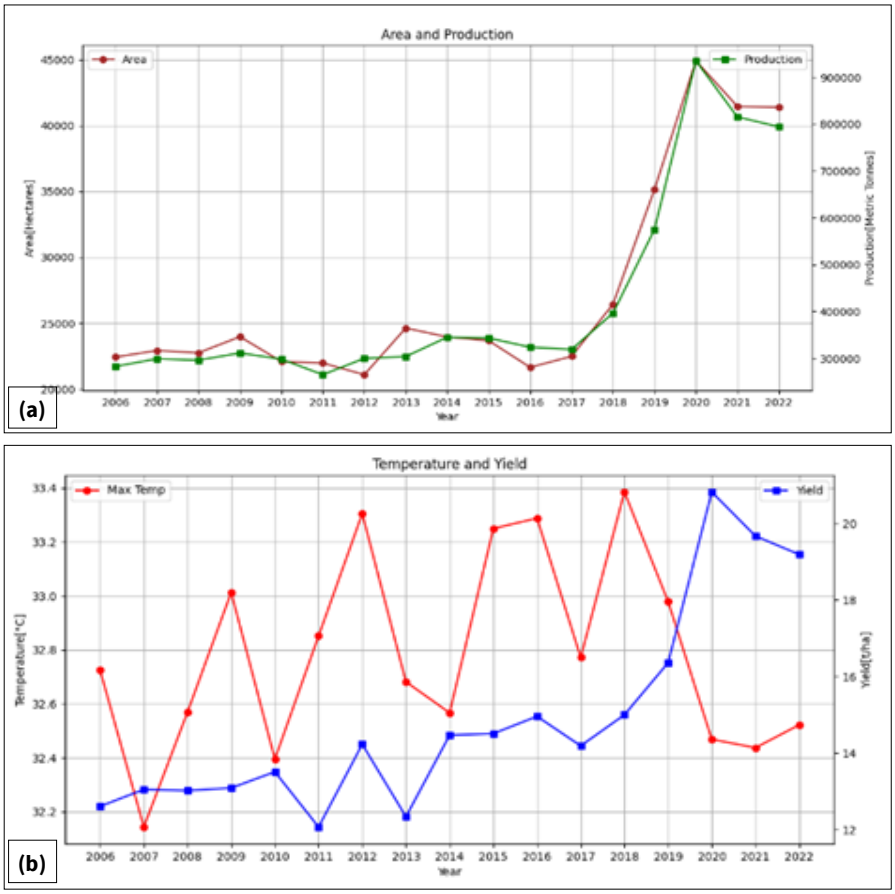

Fig. 9. (a) Area (hectares) and Production (kg/ha) of Tomato in Tamil Nadu; (b) Maximum temperature (°C) and Tomato yield (t/ha) of Tamil Nadu.

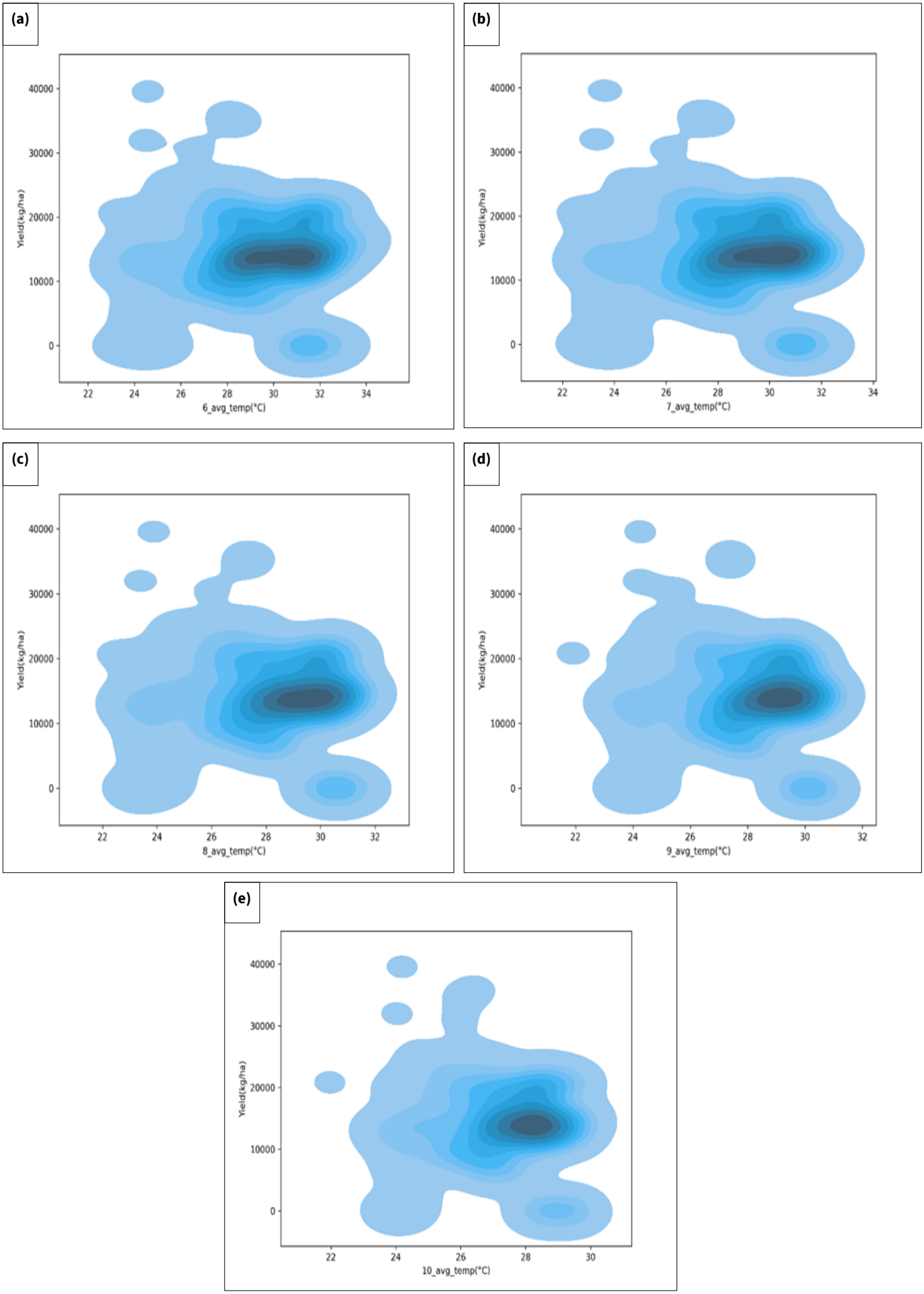

Fig. 10. Correlation between tomato productivity and monthly average temperatures. 6 - June; 7 - July; 8 - August; 9 - September; 10 - October months.

Table 1. Performance Evaluation of Model Prediction using validation scores

Model	MSE		RMSE		R^2		MAE	
	T-T	K-fold	T-T	K-fold	T-T	K-fold	T-T	K-fold
RF	7.88	9.12	2.81	2.95	0.84	0.82	1.19	1.32
XG	8.76	9.62	2.96	3.07	0.82	0.80	1.60	1.75
KNN	18.37	21.21	4.29	4.60	0.62	0.56	3.05	3.15

RF - Random Forest; XG - XGBoost; KNN - K- Nearest Neighbour; MSE - Mean Square Error; RMSE - Root Mean Square Error; R^2 - Coefficient of Determination; MAE - Mean Absolute Error; T-T - Train - Test split; K-fold - K-fold cross validation.

indicate better yield prediction by the model, followed by XG, while KNN recorded poor performance. T-T appeared to perform better than K-fold cross-validation for this specific dataset, but this may be due to overfitting. It has lower MSE, RMSE and MAE and a greater R^2 score with all prediction models.

The correlation between the predicted and actual tomato yields is depicted in Fig. 11. With occasional outliers, the RF (Fig. 11 a) model tends to be accurate, exhibiting limitations within certain yield intervals, followed by the XG model (Fig. 11 b), suggesting precise and reliable yield predictions. KNN (Fig. 11 c) showed lower performance and a weaker correlation compared with RF and XG.

Future Tomato Yield by Random Forest

Future tomato yield was projected using the Random Forest model under two climate scenarios: SSP2-4.5 and SSP5-8.5, as shown in Fig. 12 (a-b). Under SSP2-4.5, yield is projected to increase gradually by approximately 0.2 t/ha in a linear trend over the next four years. In contrast, SSP5-8.5 shows a sharper rise of 0.25 t/ha in the first two years (2023-2024), followed by a decline of 0.2 t/ha in the third year (2025), suggesting a potential negative impact of extreme heat conditions on productivity.

Discussion

Temperature requirement for tomato and its impact on crop yield

High temperatures can lead to reduced crop yield due to decreased photosynthesis. In tomato, photosynthesis occurs optimally at a temperature of around 24 °C and can be adversely affected at elevated temperatures (21). The net photosynthesis in tomato leaves increases from 18 °C to approximately 23 °C, then steadily declines up to 38 °C, with the optimal photosynthetic temperature around 24 °C under ambient CO_2 (22). The evaluation of temperature requirement was determined for five months in the study, as they are the most critical periods. The establishment stage requires a temperature of 25 °C. The ideal temperature range for the second and third months are 22 °C-26 °C and 23 °C-25 °C, respectively (23). Flowering and fruit setting stages require 18 °C-24 °C for optimum yield. The ideal temperature range for the maturity stage is 18 °C-24 °C. Temperatures exceeding 35 °C may decrease fruit set and induce flower drop (24).

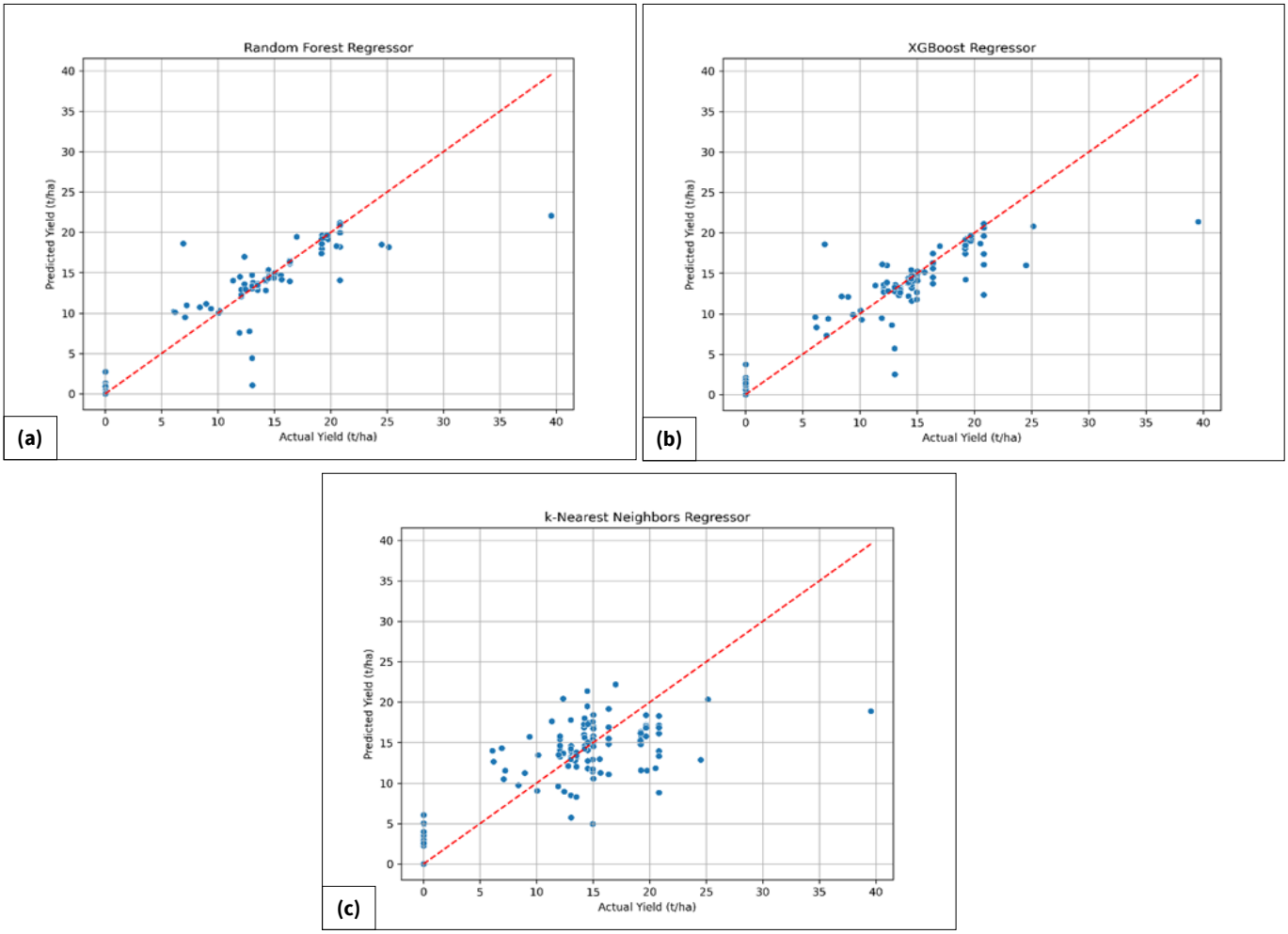

Fig. 11. Correlation between predicted and actual yield of tomato (a) Random Forest (b) XGBoost (c) K-Nearest Neighbour.

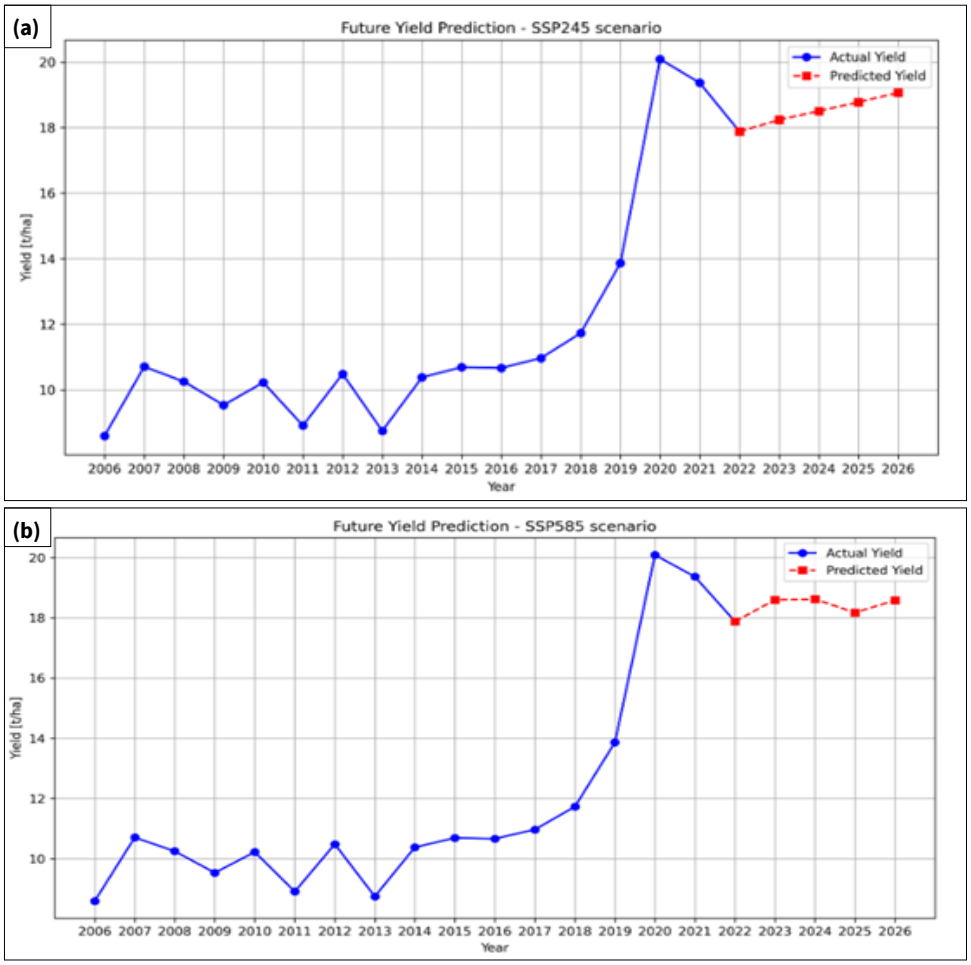

Fig. 12. Future tomato yield from 2023 to 2026 using Random Forest MLM for SSP2-4.5 (a) and SSP5-8.5 (b); SSP - Shared Socio-economic Pathway.

Temperatures above 32 °C-35 °C induce distinct heat stress symptoms, including pollen sterility, impaired pollen tube growth and flower abscission, leading to significant yield reductions (25). High temperatures also induce oxidative stress, increasing reactive oxygen species (ROS) in reproductive tissues, leading to pollen abortion and disrupted carbohydrate metabolism in anthers (26). These physiological disruptions collectively result in poor fruit set and yield loss, underscoring the sensitivity of tomato thermal extremes.

Model prediction and accuracy

In this study, MLMs such as RF, XG and KNN were considered to determine the best yield predictor of tomato for Tamil Nadu. Air Temperature was considered the most important variable in this study, along with the area of cultivation and the respective years. In accordance with the model performance evaluation, RF proved to be the best model to be recommended for developing an early crop production forecast system, followed by XGBoost which performed nearly on par with RF. The model in this study predicted an R^2 value of 0.84, whereas the most accurate predictor, the RF model, achieved low error values and a high correlation coefficient (R^2= 0.99) (27). MLM and the predictors (independent variables) are selected based on certain criteria, including the size of dataset, availability and nature of dependent variable (28). The air temperature was regarded as an important variable in this study, as crop productivity is substantially affected by temperature rise due to climate change. While the prediction results were good, incorporating more data will further increase the model's accuracy. All the performance metrics indicated that RF outperformed multiple linear regression standards and was shown to be very effective in crop yield prediction.

A study in United States found that the RMSE for RF models varied from 6-14 % of the average observed yield for wheat, maize and potato. Because of its high accuracy and precision, intuitiveness and utility in data analysis, RF served as a successful and adaptable ML technique for predicting crop yield at regional and global levels (29, 30). Meanwhile, the XGBoost regressor outperformed various other models-including Linear Regression (LR), Support Vector Regression (SVR), RF, AdaBoost (ABR), Bagging Regression (BR) and Gradient Boosting Regression (GBR)-further supporting its reliability for predictive modelling in agriculture (31).

The RF was used to simulate future tomato yield for SSP245 and SSP585 scenarios as they represent normal and extreme conditions respectively. The simulation projected that tomato yield will increase by 0.2 t/ha/year under SSP245, while under SSP585, a rise of 0.25 t/ha/year followed by a fall 0.2 t/ha/year was observed. Even though the temperatures are projected to increase in the near future, the model simulation denoted an increase in yield which may be attributed to several reasons. Climate change involves an increase in temperature along with the increasing CO_2 concentration in the atmosphere. The increased CO_2 levels can enhance the photosynthetic rate and reduce photorespiration in the C_3 plant tomato. This may compensate the detrimental effects of high temperature, leading to improved yields in short period. However, over time, the effect of CO_2 may plateau and the continued rise in

temperature may negatively affect crop yields (32).

While the model demonstrated high predictive accuracy, there are certain limitations. One potential concern is overfitting, especially given the relatively high R^2 value (0.84). Although techniques such as cross-validation were explored, overfitting remains a risk, particularly when the model is exposed to highly variable future data. Additionally, the models are sensitive to the quality and variability of input data, as only temperature was used as the predictor in the study. Other parameters, such as rainfall and humidity might also have compensated the effect of temperature (33). The exclusion of these variables, such as rainfall, humidity and soil moisture, is a limitation of this study. The absence of other agronomic and environmental variables may limit the model's accuracy across different agro-climatic zones and cropping systems. Future efforts should focus on incorporating a wider range of input variables and evaluating model performance across broader temporal and spatial scales to improve robustness and reduce prediction bias. Increase in protected cultivation of tomato, adoption of resistant varieties and other adaptation strategies may also contribute to the model's projected trend since the historic data also showed similar events of increased yield despite warmer temperatures, the model could have extrapolated it for simulation.

Conclusion

The RF model produced precise forecasts that closely resemble actual yields, making it one of the best models for predicting tomato yield for Tamil Nadu. The models were evaluated using government yield data and IMD temperature data. XGBoost also performed well, with only a marginal difference in accuracy compared to RF. These findings highlight the potential of temperature-based yield forecasting to support early warning systems, optimize planting schedules and guide government procurement and pricing strategies to stabilize markets and reduce post-harvest losses. The future yield projections indicate an initial increase in productivity by approximately 0.2 t/ha with moderate temperature rise, followed by a decline of 0.2 t/ha under higher heat stress, reinforcing the need for climate-resilient crop planning. Integrating RF models with Internet of Things-based real-time monitoring systems could enable continuous data collection and localized forecasting, aiding both farmers and policymakers in making timely, data-driven decisions. Future studies should also explore integrating additional climatic and agronomic variables to further enhance predictive accuracy and adaptability.

Acknowledgements

I extend my profound thanks to the UGC-NET JRF fellowship under the University Grants Commission, Govt. of India, New Delhi for financially aiding me to carry out this research. I extend my gratitude to the GoI - DST- CCP - NMSKCC - Centre of Excellence climate and disaster resilient agriculture for helping me conducting this research.

Authors' contributions

MC wrote the original draft; MC and MN conceptualized the study. MC, SNK and DM performed the methodology, did formal analysis and investigated the study; MN, IRC and SE reviewed and edited the article; SN and SP supervised the study.

Compliance with ethical standards

Conflict of interest: Authors do not have any conflict of interests to declare.

Ethical issues: None

References

1. Martí R, Roselló S, Cebolla-Cornejo J. Tomato as a source of carotenoids and polyphenols targeted to cancer prevention. Cancers (Basel). 2016;8(6):58. https://doi.org/10.3390/cancers8060058
2. Kheyrodin H, Kheyrodin S. Importance of the tomato as such as medical plant. Int J Adv Res Biol Sci. 2017;4(4):106-15. https://doi.org/10.22192/ijarbs.2017.04.04.014
3. Mariyappan D, Srividhya S, Vennila MA, Sasikumar K, Senthilkumar T. Study the performance of high yielding tomato hybrids in Dharmapuri District, Tamil Nadu, India. J Adv Biol Biotechnol. 2024;27:1-9. https://doi.org/10.9734/jabb/2024/v27i7993
4. Malathi G, Kohila P. Evaluation of tomato hybrids in Salem district of Tamil Nadu. J Krishi Vigyan. 2021;10(1):328–31. https://doi.org/10.5958/2349-4433.2021.00115.X
5. Kurina AB, Solovieva AE, Khrapalova IA, Artemyeva AM. Biochemical composition of tomato fruits of various colors. Vavilov J Genet Breed. 2021;25(5):514–20. https://doi.org/10.18699/VJ21.058
6. Quinet M, Angosto T, Yuste-Lisbona FJ, Blanchard-Gros R, Bigot S, Martinez JP, et al. Tomato fruit development and metabolism. Front Plant Sci. 2019;10:1554. https://doi.org/10.3389/fpls.2019.01554
7. Camejo D, Rodríguez P, Morales MA, Dell'Amico JM, Torrecillas A, Alarcón JJ. High temperature effects on photosynthetic activity of two tomato cultivars with different heat susceptibility. J Plant Physiol. 2005;162(3):281–9. https://doi.org/10.1016/j.jplph.2004.07.014
8. Alsamir M, Ahmad NM, Keitel C, Mahmood T, Trethowan R. Identification of high-temperature tolerant and agronomically viable tomato (*Solanum lycopersicum*) genotypes from a diverse germplasm collection. Adv Crop Sci Technol. 2017;5(110):1–6. https://doi.org/10.4172/2329-8863.1000299
9. Hazra P, Ansary SH, Dutta AK, Balacheva E, Atanassova B. Breeding tomato tolerant to high temperature stress. Acta Hortic. 2009;830:241–8. https://doi.org/10.17660/ActaHortic.2009.830.33
10. Berry SZ, Uddin MR. Effect of high temperature on fruit set in tomato cultivars and selected germplasm. HortScience. 1988;23(3):606–8. https://doi.org/10.21273/HORTSCI.23.3.606
11. Ozores-Hampton M, Kiran F, McAvoy G. Blossom drop, reduced fruit set and post-pollination disorders in tomato. Electronic Data Information Source. 2012;2012(7). https://doi.org/10.32473/edis-hs1195-2012
12. Yadav MR, Choudhary M, Singh J, Lal MK, Jha PK, Udawat P, et al. Impacts, tolerance, adaptation and mitigation of heat stress on wheat under changing climates. Int J Mol Sci. 2022;23(5):2838. https://doi.org/10.3390/ijms23052838
13. Javadinejad S, Eslamian S, Ostad-Ali-Askari K. The analysis of the most important climatic parameters affecting performance of crop variability in a changing climate. Int J Hydrol Sci Technol. 2021;11(1):1–25. https://doi.org/10.1504/IJHST.2021.112651
14. Beillouin D, Schauberger B, Bastos A, Ciais P, Makowski D. Impact of extreme weather conditions on European crop production in 2018. Philos Trans R Soc Lond B Biol Sci. 2020;375(1810):20190510. https://doi.org/10.1098/rstb.2019.0510

15. Reddy DJ, Kumar MR. Crop yield prediction using machine learning algorithm. In: 2021 5th Int Conf Intell Comput Control Syst (ICICCS). IEEE. 2021:1466–70. https://doi.org/10.1109/ICICCS51141.2021.9432236

16. Hancock PA, Hutchinson MF. Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. Environ Model Softw. 2006;21(12):1684–94. https://doi.org/10.1016/j.envsoft.2005.08.005

17. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–40. https://doi.org/10.1007/BF00058655

18. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324

19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016:785–94. https://doi.org/10.1145/2939672.2939785

20. Mayer DG, Butler DG. Statistical validation. Ecol Model. 1993;68(1–2):21–32. https://doi.org/10.1016/0304-3800(93)90105-2

21. Ku SB, Hunt LA. Effects of temperature on the photosynthesis–irradiance response curves of newly matured leaves of alfalfa. Can J Bot. 1977;55(8):872–9. https://doi.org/10.1139/b77-106

22. Boote KJ, Rybak MR, Scholberg JM, Jones JW. Improving the CROPGRO-tomato model for predicting growth and yield response to temperature. HortScience. 2012;47(8):1038–49. https://doi.org/10.21273/HORTSCI.47.8.1038

23. Vijaykumar A, Beena R. Response of tomato quality and yield to elevated temperatures under controlled environment. Int J Environ Clim Change. 2023;13(12):256–71. https://doi.org/10.9734/ijecc/2023/v13i123682

24. Kuradusenge M, Hitimana E, Hanyurwimfura D, Rukundo P, Mtonga K, Mukasine A, et al. Crop yield prediction using machine learning models: case of Irish potato and maize. Agriculture. 2023;13(1):225. https://doi.org/10.3390/agriculture13010225

25. Müller F, Xu J, Kristensen L, Wolters-Arts M, de Groot PF, Jansma SY, et al. High-temperature-induced defects in tomato (*Solanum lycopersicum*) anther and pollen development are associated with reduced expression of B-class floral patterning genes. PLoS One. 2016;11(12):e0167614. https://doi.org/10.1371/journal.pone.0167614

26. Huang X, Xiao N, Xie Y, Xu C. ROS burst prolongs transcriptional condensation to slow shoot apical meristem maturation and achieve heat-stress resilience in tomato. Dev Cell. 2025. https://doi.org/10.1016/j.devcel.2025.03.007

27. Haque E, Tabassum S, Hossain E. A comparative analysis of deep neural networks for hourly temperature forecasting. IEEE Access. 2021;9:160646–60. https://doi.org/10.1109/ACCESS.2021.3131533

28. Crane-Droesch A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ Res Lett. 2018;13(11):114003. https://doi.org/10.1088/1748-9326/aae159

29. Everingham Y, Sexton J, Skocaj D, Inman-Bamber G. Accurate prediction of sugarcane yield using a random forest algorithm. Agron Sustain Dev. 2016;36:64. https://doi.org/10.1007/s13593-016-0364-z

30. Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE, et al. Random forests for global and regional crop yield predictions. PLoS One. 2016;11(6):e0156571. https://doi.org/10.1371/journal.pone.0156571

31. Ge J, Zhao L, Yu Z, Liu H, Zhang L, Gong X, et al. Prediction of greenhouse tomato crop evapotranspiration using XGBoost machine learning model. Plants. 2022;11(15):1923. https://doi.org/10.3390/plants11151923

32. Ayankojo IT, Morgan KT. Increasing air temperatures and its effects on growth and productivity of tomato in South Florida. Plants. 2020;9(9):1245. https://doi.org/10.3390/plants9091245

33. Nonhebel S. Effects of temperature rise and increase in CO_2 concentration on simulated wheat yields in Europe. Clim Change. 1996;34(1):73–90. https://doi.org/10.1007/BF00139254

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

Publisher information: Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.